J Clin Epidemiol Vol. 44, No. 2, pp. 103–107, 1991
Printed in Great Britain

0895-4356/91 $3.00 + 0.00
Pergamon Press plc

Editorial

META-META-ANALYSIS: UNANSWERED QUESTIONS ABOUT AGGREGATING DATA

WALTER O. SPITZER*

Department of Epidemiology and Biostatistics, McGill University, 1020 Pine Avenue West, Montreal, Quebec, Canada H3A 1A2

(Received for publication 30 October 1990)

Some days I ask myself if investigators still do their own trials, with admissible standards and adequate power. It seems like the most frequent prelude to a discussion on efficacy or safety of an intervention is, "... We did a meta-analysis of studies on ...". Does anyone do straightforward unhyphenated analyses any more?

In this issue of the **Journal** (pp. 127–139), Fleiss and Gross review and reassess meta-analysis and report an illustrative case study of the method focusing on the association between exposure to environmental tobacco smoke (ETS) and lung cancer. The article refers to a succinct and useful definition of meta-analysis offered by Huque [1] "... a statistical analysis which combines or integrates the results of several independent clinical trials considered by the analyst to be 'combinable'". I also take Huque's characterization as the working definition for this editorial comment and join Fleiss and Gross in highlighting the fact that meta-analysis' key and almost exclusive application to date has been in the integration of data from clinical trials. I would add that a distinctive characteristic of the strategy is the derivation of a single quantitative estimate of effect of an intervention or a risk factor. Fleiss and Gross touch on most of the main issues, including consensus and controversy. I will not repeat or summarize their elegant review which enumerates many of the unanswered questions about meta-analysis of experimental trials. But it deserves emphasis that the main unresolved challenges are to settle on universal widely acceptable criteria for exclusion of trials and the development of a reproducible, valid and accepted weighting index that would enable analysts to invoke the quality of the research into a final result. I will confine my comments mostly to the application of meta-analysis in aggregation of non-experimental (observational) studies. The controversies surrounding meta-analysis of experimental trials are equally relevant to non-experimental studies which are usually epidemiological. But there are additional unanswered questions. Fleiss and Gross ask, "... can meta-analytic techniques be applied in the analysis of other kinds of data such as those that arise in cohort and case–control studies found in epidemiology? Their answer is a "guarded 'yes'". I do not know whether the question can be answered at all. The illustrative meta-analytic project reported by them goes a long way in avoiding potential pitfalls. However, some problems are not fully addressed either in the execution of the study nor in the discussion about such applications in conventional epidemiology. There are many difficulties that have not been surmounted yet, either theoretically or empirically. When phrasing the following 13 unanswered questions, I have used words such as "merge", "combine", "integrate" or "put together", in respect to data from different case series, different series of reference groups, different cohorts, etc. I have done so recognizing that one usually

*Reprint requests should be addressed to W. O. Spitzer at the above address.

2023381903

Source: https://www.industrydocuments.ucsf.edu/docs/mhyx0000

combines intrastudy *differences*, not intrastudy disease or exposure rates. I ask the reader's indulgence for having avoided repetitive dense statistical technical terminology in deference to readable English prose. These then, are the questions:

(1) Operationally, what are the "stringent conditions" (Fleiss and Gross' phrase) under which both case–control studies and cohort studies may be included in one single meta-analysis? Should such analyses ever be done without access to the raw data of the component studies?

(2) When is it permissible to combine different types of cohort? For instance, for both exposed cohorts and comparison cohorts should one integrate data from a fixed cohort with an open one?

(3) Is it permissible to integrate exposed patients sampled from hospitals with those from primary care settings?

(4) For reference cohorts, *not exposed* to an intervention or risk factor, other questions arise. For example,

  (a) Is a comparison cohort from Sweden combinable with one from Italy or Japan?

  (b) Are cohorts taken from occupational sampling frames sufficiently similar to those from the corresponding general population (or another geographically-defined one) to put them together?

  (c) How separate in time must the accrual or demarcation of unexposed cohorts become to be ineligible for aggregation? (The question is also pertinent for exposed cohorts.)

Turning now to case–control designs:

(5) Is it admissible to merge hospital-based with population-based case groups? Or in Miettinen's terms, can two or more case series be combined if they are not representative of the same type of base experience? [2].

(6) Conceptually, and in execution, is a nested case–control study similar enough to a conventional case–control study for both to be included in the same meta-analysis?

(7) When there are two or more control groups in a case–control study does one merge all the control groups? If not, what criteria must one use to exclude any control group from the meta-analysis? There is no parallel between multiple arms defined by exposure in a randomized controlled trial and multiple reference samples demarcated by outcome in a case–control study.

(8) Should control groups assembled by matching be combined with independent samples of referenced populations?

(9) What constitutes "proper control or adjustment for the biases that frequently occur in epidemiological studies?" [3]. Case–control studies are the designs in common use most vulnerable to bias [4]. I believe that the difficulties in minimizing bias acceptably, alone, could vitiate the validity of meta-analyses of case–control studies. Bias is barely manageable (and seldom managed well) even in single case–control studies. As pointed out by Fleiss and Gross [3], as well as Letzl earlier [5], misclassification bias for exposure is a particularly thorny problem.

(10) Are data provided by proxy informants similar enough to data from respondents to be considered equivalent?

(11) Should one include case–control studies in which data-gatherers were unblinded with blinded studies in one meta-analysis? (Should one do so in cohort research?)

(12) How homogeneous must the outcome be? For instance, can one pool data from a study that ascertained "all cancers of the lung", with one that did so only for "oat cell Ca", or only "adenocarcinoma"?

(13) How do we interpret values and confidence intervals of single estimates derived with meta-analysis? Consider a report of a single study comparing two fixed cohorts of 2500 persons. The relative risk (RR) for the association of the putative risk factor with the incidence of a well-defined, hard, but relatively uncommon outcome is 1.45. The 95% confidence interval is 1.02–2.07 and $p = 0.04$. In a hypothetical meta-analysis of five other follow-up studies with 500 persons per cohort (two in each study assessing the same risk factor and outcome) the result is an identical relative risk, the same confidence interval and an

2023381904

Source: https://www.industrydocuments.ucsf.edu/docs/mhyx0000

identical p value. In the second scenario the relative risks for the five component studies were 3.0 ($p = 0.013$), 0.6 ($p = 0.48$), 1.1 ($p = 0.80$), 1.5 ($p = 0.15$), 1.1 ($p = 0.83$). Do both sets of statistics mean the same thing? I remind the reader that it is not standard practice to incorporate interstudy variation with one's meta-analysis. Should one do so? Beyond problems of multiple comparisons which we usually recognize and correct, might there be a problem of "multiple combinations"?

As summarized by Fleiss and Gross, the indications for a properly conducted meta-analysis are,

(a) to increase statistical power,
(b) to deal with controversy when individual studies disagree,
(c) to improve estimates of size of effect and
(d) to answer new questions not previously posed in component studies [3]. But is it always necessary, or justified, to pursue a single estimate with its related probability qualifiers to derive conclusions from a series of research projects?

An alternative to both meta-analysis on one hand, and traditional (often haphazard) reviews is an approach proposed by Slavin and designated *best-evidence synthesis*. This approach considers that the "best evidence" in any field comes from studies having the highest internal and external validity, that use well-specified, defined, explicit *a priori* inclusion and exclusion criteria and favour size-effect data to statistical significance alone when interpreting the literature reviewed. Such syntheses emphasize numeric findings but the conclusions need not depend on a single estimation nor on statistical significance [6]. In common with properly executed meta-analyses best-evidence syntheses cannot evade the difficult challenge of deciding what to exclude and how to document the exclusions. Getting around "publication bias" [7] which means finding and judging unpublished work is particularly daunting.

What is attractive about best-evidence synthesis is that it liberates the analyst from the apparent obsession which meta-analysts have to calculate a single estimate as a necessary intermediate step to reach an opinion about an association, an effect or a casual relationship. Nevertheless, best-evidence synthesis does not exonerate the analyst from the highest attainable rigour in setting forth a protocol for the synthesis in advance. The protocol must then be followed when deciding which component studies reach a predetermined level of scientific admissibility, in establishing exclusion criteria, when implementing methods to document excluded research, when developing valid weighting schemes for the quality of papers, and when formulating predetermined explicit rules for judging effect size. The foregoing list of specifications for a best-evidence synthesis protocol is not exhaustive. I am of the opinion that even more rigour is required of the meta-analyst.

Turning now to Fleiss and Gross' illustrative case study that re-examines the association between exposure to environmental tobacco smoke and lung cancer [3], I shall highlight some of its features:

(i) They excluded non-American studies, a sensible move, given the unresolved methodological problems of pooling people of very different ethnic nature and of different culturally-determined views of smoking. The exclusion is particularly appropriate if the intention was to make inferences chiefly about the American population.
(ii) The meta-analysis incorporates one cohort study and eight case–control studies.
(iii) Hospital-based and population-based groups (both cases and controls) appear to have been considered equivalent. It is probably valid to have done so *when the purpose was to test the null hypothesis* of no association. Were one to attempt non-null inferences, it would have been a mistake to consider them equivalent.
(iv) Matched and unmatched controls were incorporated in a similar way.
(v) The overall analysis does not seem to have been adjusted by blindness status of data-gatherers nor by the extent that proxies or respondents provided information on exposure.
(vi) Outcomes were somewhat heterogeneous. Consequently, matched groups might have been different.

Lastly, they report,

(vii) "... we did not develop *a priori*, a set of procedures for the unbiased measurement of a study's quality ...".

2023381905

Source: https://www.industrydocuments.ucsf.edu/docs/mhyx0000

The seven comments are not made to rebuke Fleiss and Gross' approach but to underscore the enormous difficulties that they and anyone else unavoidably confront attempting a meta-analysis of non-experimental epidemiological studies.

The single estimate they report for the nine studies is 1.12 (CI 95% 0.95–1.30), $\chi^2 = 1.88$ (1 *df*). I consider the result and the resulting conclusion plausible and I completely agree with their comment. "The fact that no significant association was found neither vindicates nor condemns the meta-analysis of epidemiological studies" [3].

My colleagues and I, in a Working Group on Passive Smoking that reported early last year [8], also examined the association between exposure to environmental tobacco smoke (ETS) and lung cancer. We used Slavin's method of best-evidence synthesis rather than meta-analysis. We also included the world literature, not just U.S. studies. It is instructive to compare the two sets of conclusions on the association:

Fleiss and Gross: "... there is no convincing scientific evidence from the epidemiologic literature of an association between exposure to ETS and the risk of lung cancer in the United States" [3].

Spitzer et al.: "The weight of evidence is compatible with a positive association between residential exposure to environmental tobacco smoke (primarily from spousal smoking) and the risk of lung cancer". "There is no evidence for an association between non-residential exposure to ETS and any form of cancer" [8].

The two conclusions are not identical. But they are not directly contradictory nor mutually exclusive. Moreover, given the restriction of the meta-analysis to U.S. studies and the inclusion of admissible studies from anywhere in the world for the best-evidence synthesis the compatibility of the "verdicts" tends to mutually buttress their validity. Admittedly, the language of the discussion of the Fleiss and Gross paper [3] favours a *non-association* interpretation while the Spitzer group's comments do the opposite [8]. For example, contrast these phrases: "... the safest conclusion from the present meta-analysis is a negative one" [3]; "The preponderance of positive studies is consistent with a causal relationship betwen exposure to ETS and lung cancer" [8]. But it is obvious that there is a lot of common ground between the conclusions of two different methodological approaches to the quantitative evidence available on the subject.

A more general conclusion of Fleiss and Gross is important: "Meta-analyses, when properly performed, can be used effectively in both clinical trials and epidemiological studies ...". In today's state of science I accept the conclusion guardedly and warily, *for clinical trials only* despite many unresolved controversies about what "properly performed" means in the method. In my opinion, however, the unanswered questions about meta-analysis in non-experimental epidemiological studies do not yet warrant widespread application except as methodological research. I view Fleiss and Gross' analysis as a courageous, honest, trailblazing early step in the development of the method. Their own caveats throughout their article lend support to my opinion and to their scientific integrity.

In the near future we need a level of international consensus about the methods of meta-analysis as high as that which prevails for unmeta-analyzed single randomized controlled trials. Perhaps a "summit" should be called for. I would find it difficult to ignore a general agreement on criteria for scientific admissibility of meta-analytic studies (both clinical trials and non-experimental studies) if it were endorsed by a group including, say, Armitage, Breslow, Cole, Day, Detsky, Gross, Feinstein, Fleiss, Meier, Miettinen, R. Peto, Sackett, Schwartz, Uberla, Vessey, Walter and Zelen. Minor miracles still happen occasionally.

Finally, let's abolish the verb "to meta-analyze" as a substitute for "to review", "to synthesize", "to interpret" or even "to read". Careless use of the technically-specific term does not do justice to the unfulfilled promise of meta-analysis, nor to the painstaking work of many colleagues who pursue excellence as they attempt to deliver the promise.

Acknowledgement—This study was supported by the National Health Resources Development Programme, Health and Welfare, Canada.

REFERENCES

1. Huque MF. Experiences with meta-analysis in NDA submissions. Proc Biopharmaceutical Section of the American Statistical Association 1988; 2: 28–33.
2. Miettinen OS. The "case-control" study: valid selection of subjects. J Chron Dis 1985; 38: 543–548.
3. Fleiss JL, Gross AJ. Meta-analysis in epidemiology, with special reference to studies of the association between exposure to environmental tobacco smoke

2023381906

Source: https://www.industrydocuments.ucsf.edu/docs/mhyx0000

and lung cancer: a critique. **J Clin Epidemiol** 1991; 44: 127–139.

4. Ibrahim MA, Spitzer WO, Eds. The case–control study: consensus and controversy. **J Chron Dis** 1979; 32: 1–90.
5. Letzel H, Blummer E, Uberla K. Meta-analyses on passive smoking and lung cancer: Effects of study selection and misclassification of exposure. **Environ Technol Lett** 1988; 9: 491–500.
6. Slavin RE. Best-evidence synthesis: an alternative to meta-analytic and tradiational reviews. **Educ Res** 1986; 15: 5–11.
7. Vandenbroucke JP. Passive smoking and lung cancer: a publication bias? **Br Med J** 1988; 296: 391–392.
8. Spitzer WO, Lawrence V, Dales R *et al*. Links between passive smoking and disease: A best-evidence synthesis. **Clin Invest Med** 1990; 13: 17–42.

2023381907

Source: https://www.industrydocuments.ucsf.edu/docs/mhyx0000